

# Nonparametric and Varying Coefficient Modal Regression

Weixin Yao <sup>\*</sup> and Sijia Xiang <sup>†</sup>

## SUMMARY

In this article, we propose a new nonparametric data analysis tool, which we call *nonparametric modal regression*, to investigate the relationship among interested variables based on estimating the mode of the conditional density of a response variable  $Y$  given predictors  $X$ . The nonparametric modal regression is distinguished from the conventional nonparametric regression in that, instead of the conditional average or median, it uses the “most likely” conditional values to measures the center. Better prediction performance and robustness are two important characteristics of nonparametric modal regression compared to traditional nonparametric mean regression and nonparametric median regression. We propose to use local polynomial regression to estimate the nonparametric modal regression. The asymptotic properties of the resulting estimator are investigated. To broaden the applicability of the nonparametric modal regression to high dimensional data or functional/longitudinal data, we further develop a nonparametric varying coefficient modal regression. A Monte Carlo

---

<sup>\*</sup>Department of Statistics, University of California, Riverside, California, U.S.A. E-mail: weixin.yao@ucr.edu. Yao’s research is supported by NSF grant DMS-1461677.

<sup>†</sup>School of Mathematics and Statistics, Zhejiang University of Finance and Economics. E-mail: sjxiang@zufe.edu.cn. Xiang’s research is supported by Zhejiang Provincial NSF of China grant LQ16A010002.

simulation study and an analysis of health care expenditure data demonstrate some superior performance of the proposed nonparametric modal regression model to the traditional nonparametric mean regression and nonparametric median regression in terms of the prediction performance.

*Some key words:* EM algorithm; Local polynomial regression; Modal regression; Mode; Robust.

## 1. INTRODUCTION

Suppose  $\{(x_i, y_i), i = 1, \dots, n\}$  is a random sample, where  $x_i$  is a  $p$ -dimensional column vector, and  $f(y | x)$  is the conditional density function of  $Y$  given  $x_i$ . For the conventional regression models, the mean of  $f(y | x)$  is usually used to investigate the relationship between  $Y$  and  $X$ . When the distribution is highly skewed, it is well known that the mode provides a more meaningful location estimator than the mean. Several authors have made efforts to identify the modes of population distributions. See, for example, Scott (1992); Friedman and Fisher (1999); Chaudhuri and Marron (1999); Hall, Minnotte, and Zhang (2004); Ray and Lindsay (2005); Yao and Lindsay (2009). Recently, Lee (1989, 1993), Lee and Kim (1998), Kemp and Santos Silva (2012), and Yao and Li (2014) successfully applied the mode idea to linear regression and proposed the linear modal regression which assumes that the mode of  $f(y | x)$ , denoted by  $\text{Mode}(y | x)$ , is a linear function of  $x$ . Better prediction performance and robustness are two important characteristics of modal regression compared to the traditional mean regressions. Please see Kemp and Santos Silva (2012) and Yao and Li (2014) for more discussions about the advantage of modal regression as a promising alternative regression tool to traditional regression models. However, in practice, the strong parametric assumption about  $\text{Mode}(y | x)$  might not hold and

thus the corresponding inference might be misleading. Therefore, it is desirable to develop some estimation procedure to relax the parametric model assumption about  $\text{Mode}(y \mid x)$ .

In this article, we propose a nonparametric modal regression model that aims to estimate the *mode* of  $f(y \mid x)$  for any given  $x$  without assuming any parametric model. Instead of the conditional average used by the traditional regression methods, modal regression uses the “most probable” conditional values to measure the center. We propose to use local polynomial regression to estimate the nonparametric modal regression. Sampling properties of the proposed estimates are systematically studied. A modal expectation-maximization (EM) algorithm is also developed for the proposed models. To broaden the applicability of the nonparametric modal regression, we further develop a nonparametric varying coefficient modal regression. A Monte Carlo simulation study and an analysis of health care expenditure data demonstrate some superior performance of the proposed nonparametric modal regression to the traditional nonparametric mean regression.

The rest of this article is organized as follows. In Section 2, we introduce the new nonparametric modal regression model and the estimation procedure based on local polynomial regression. The asymptotic properties of the resulting estimator are also provided. In Section 3, we propose a nonparametric varying coefficient modal regression. A Monte Carlo simulation study and a real data application are conducted in Section 4 to illustrate the proposed models. We conclude our article by some discussions in Section 5.

## 2. NONPARAMETRIC MODAL REGRESSION

In this section, we will introduce the nonparametric modal regression model and the estimation procedure based on local polynomial regression. An EM type algorithm is

proposed to estimate the unknown modal parameters. In addition, we will also study the asymptotic properties of the proposed estimator.

### 2.1. Model introduction

Suppose that  $(x_1, y_1), \dots, (x_n, y_n)$  are an independent and identically distributed random sample from  $f(x, y)$ . The modal regression is defined as

$$m(x) = \text{Mode}(y \mid x) = \arg \max_y f(y \mid x), \quad (2.1)$$

where  $m(\cdot)$  is an unknown nonparametric smoothing function to be estimated. For simplicity of explanation, we assume that  $x$  is a scalar but the proposed model can be extended to the multivariate predictor  $x$ . However, such extension is less desirable due to the “curse of dimensionality”.

Let  $\epsilon = y - m(x)$ . Denote by  $g(\epsilon \mid x)$  the conditional density of  $\epsilon$  given  $X = x$ . Based on the model assumption (2.1), one can know that  $g(\epsilon \mid x)$  is maximized at 0 for any  $x$ . If  $g(\epsilon \mid x)$  is symmetric about 0, then  $m(x)$  is the same as the conventional regression function  $E(Y \mid X = x)$ . In this article, we propose an estimation procedure for the nonparametric modal regression  $m(x)$ .

Since  $f(y \mid x) = f(x, y)/f(x)$ , finding the mode of  $f(y \mid x)$  is equivalent to finding the mode of  $f(x, y)$  with  $x$  fixed. Suppose  $f(x, y)$  is estimated by the kernel density estimator, i.e.,

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_i - x) \phi_{h_2}(y_i - y),$$

where  $K_h(x) = h^{-1}K(x/h)$  and  $\phi_h(t) = h^{-1}\phi(t/h)$  are the symmetric kernel functions and  $(h_1, h_2)$  are the bandwidths.

A natural estimation procedure is to estimate  $m(x_0)$  by

$$\hat{m}(x_0) = \arg \max_y \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_i - x_0) \phi_{h_2}(y_i - y). \quad (2.2)$$

Note that the modal regression (2.2) only uses one term (the intercept) for the conditional mode, like the local constant estimator (Nadaraya, 1964; Watson, 1964). It is known that the local linear estimator is superior to the local constant one (Fan and Gijbels, 1996), and so we may want to extend the idea of local constant modal regression (2.2) to the local linear case, or more generally, the local polynomial case. For  $x$  in a neighborhood of  $x_0$ , we approximate

$$m(x) \approx \sum_{j=0}^p \frac{m^{(j)}(x_0)}{j!} (x - x_0)^j \equiv \sum_{j=0}^p \beta_j (x - x_0)^j,$$

where  $\beta_j = m^{(j)}(x_0)/j!$ . Our local polynomial modal regression (LPMR) estimation procedure is to *maximize* over  $\boldsymbol{\theta} = (\beta_0, \dots, \beta_p)$

$$\ell(\boldsymbol{\theta}) \equiv \frac{1}{n} \sum_{i=1}^n K_{h_1}(x_i - x_0) \phi_{h_2} \left( y_i - \sum_{j=0}^p \beta_j (x_i - x_0)^j \right). \quad (2.3)$$

For ease of computation, we use the standard normal density for  $\phi(t)$  throughout this article. See (2.4) below. Denote the maximizer of  $\ell(\boldsymbol{\theta})$  as  $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \dots, \hat{\beta}_p)$ . Then the estimator of the  $v$ -th derivative of  $m(x)$ ,  $m^{(v)}(x)$ , will be  $\hat{m}_v(x_0) = v! \hat{\beta}_v$ , for  $v = 0, \dots, p$ . Specifically, when  $p = 1$  and  $v = 0$ , we refer to this method as the local linear modal regression (LLMR).

Note that Yao et al. (2012) used an objective function similar to (2.3) to provide an adaptive robust nonparametric regression estimate. However, the model setting and the assumptions on tuning parameters in this article are completely different

from theirs. Although Yao et al. (2012) also named their method modal regression, similar to the traditional robust regression, they assume that the error distribution is symmetric about 0 in order to get a consistent estimate. Therefore, Yao et al. (2012) still focused on mean regression, even though they motivated their estimation procedure from a modal regression point of view. The nonparametric modal regression we proposed in this article, however, allows the error distribution to be skewed or even depend on  $x$  and is truly targeting the conditional mode of  $f(y \mid x)$ . In addition, in Yao et al. (2012),  $h_2$  is a fixed value and does not depend on  $n$ . In this article, we assume that  $h_2$  goes to 0 in order to get the consistent modal regression estimate under very mild assumption of the error distribution. Moreover, we show in Section 2.3 that the asymptotic results, such as convergence rates, of the proposed estimate are completely different from Yao et al. (2012).

## 2.2. Computation algorithm

Note that (2.3) does not have an explicit solution. Similar to Yao et al. (2012), we can use an EM algorithm to maximize (2.3) since it has a mixture type form. For easy reference, we also describe the algorithm below.

Let  $\boldsymbol{\theta}^{(0)}$  be the initial value. Starting with  $k = 0$ :

*E-Step:* Update  $\pi(j \mid \boldsymbol{\theta}^{(k)})$

$$\pi(j \mid \boldsymbol{\theta}^{(k)}) = \frac{K_{h_1}(x_j - x_0)\phi_{h_2} \left\{ y_j - \sum_{l=0}^p \beta_l^{(k)}(x_j - x_0)^l \right\}}{\sum_{i=1}^n \left[ K_{h_1}(x_i - x_0)\phi_{h_2} \left\{ y_i - \sum_{l=0}^p \beta_l^{(k)}(x_i - x_0)^l \right\} \right]}, j = 1, \dots, n.$$

*M-Step:* Update  $\boldsymbol{\theta}^{(k+1)}$

$$\begin{aligned} \boldsymbol{\theta}^{(k+1)} &= \arg \max_{\boldsymbol{\theta}} \sum_{j=1}^n \left[ \pi(j \mid \boldsymbol{\theta}^{(k)}) \log \phi_{h_2} \left\{ y_j - \sum_{l=0}^p \beta_l(x_j - x_0)^l \right\} \right] \\ &= (\mathbf{X}^T \mathbf{W}_k \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_k \mathbf{Y}, \end{aligned} \tag{2.4}$$

since  $\phi(\cdot)$  is the density function of a standard normal distribution. Here  $\mathbf{X} = (\mathbf{x}_1^*, \dots, \mathbf{x}_n^*)^T$  with  $\mathbf{x}_i^* = \{1, x_i - x_0, \dots, (x_i - x_0)^p\}^T$ ,  $\mathbf{W}_k$  is an  $n \times n$  diagonal matrix with diagonal elements  $\pi(j \mid \boldsymbol{\theta}^{(k)})$ s, and  $\mathbf{Y} = (y_1, \dots, y_n)^T$ .

Similar to the usual EM algorithm, the value the algorithm converged to might rely on the starting values, and it is not certain that the algorithm converges to the global optimum. Thus, initiating the algorithm from different starting values and then choosing the best local optimal solution is vital.

### 2.3. Theoretical properties

We first establish the convergence rate of the LPMR estimator in the following theorem, whose proof can be found in the Appendix.

**Theorem 2.1.** *Under the regularity conditions (A1)—(A4) in the Appendix, if the bandwidths  $h_1$  and  $h_2$  go to 0 such that  $nh_1h_2^5 \rightarrow \infty$  and  $h_1^{p+1}/h_2 \rightarrow 0$ , there exists a consistent local maximizer  $\hat{\boldsymbol{\theta}}$  of (2.3) such that*

$$|h_1^v \{\hat{m}_v(x_0) - m^{(v)}(x_0)\}| = O_p \{(nh_1h_2^3)^{-1/2} + h_1^{p+1} + h_2^2\}, \quad v = 0, 1, \dots, p,$$

where  $\hat{m}_v(x_0) = v!\hat{\beta}_v$  is the estimate of  $m^{(v)}(x_0)$  and  $m^{(v)}(x_0)$  is the  $v^{th}$  derivative of  $m(x)$  at  $x_0$ .

The proof of Theorem 2.1 is given in the Appendix. To derive the asymptotic bias and variance of the LPMR estimator, we need the following notations. The moments of  $K$  and  $K^2$  are denoted respectively by

$$\mu_j = \int t^j K(t) dt \quad \text{and} \quad \nu_j = \int t^j K^2(t) dt.$$

Let  $S$ ,  $\tilde{S}$ , and  $S^*$  be  $(p+1) \times (p+1)$  matrix with  $(j, l)$ -element  $\mu_{j+l-2}$ ,  $\mu_{j+l-1}$ , and  $\nu_{j+l-2}$ , respectively, and  $c_p$ ,  $\tilde{c}_p$ , and  $c_p^*$  be  $p \times 1$  vector with  $j$ -th element  $\mu_{p+j}$ ,  $\mu_{p+j+1}$ ,

and  $\mu_{j-1}$ , respectively. Furthermore, let  $e_{v+1} = (0, \dots, 0, 1, 0, \dots, 0)^T$  be a  $p \times 1$  vector with 1 in the  $(v+1)^{th}$  position.

**Theorem 2.2.** *Under the regularity conditions (A1)–(A4) in the Appendix, if the bandwidths  $h_1$  and  $h_2$  go to 0 such that  $nh_1^3 h_2^5 \rightarrow \infty$  and  $h_1^{p+1}/h_2 \rightarrow 0$ , the asymptotic variance of  $\hat{m}_v(x_0)$ , given in Theorem 2.1, is*

$$\text{Var}\{\hat{m}_v(x_0)\} = e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1} \frac{v!^2 g(0 | x_0) \tilde{\nu}}{f(x_0) g''(0 | x_0)^2 n h_1^{1+2v} h_2^3} \{1 + o_p(1)\},$$

where  $\tilde{\nu} = \int t^2 \phi^2(t) dt$ . The asymptotic bias of  $\hat{m}_v(x_0)$ , denoted by  $b_v(x_0)$ , for  $p - v$  odd is given by

$$b_v(x_0) = e_{v+1}^T S^{-1} \left\{ h_1^{p+1-v} \frac{v!}{(p+1)!} m^{(p+1)}(x_0) c_p - \frac{g'''(0 | x_0) v! h_2^2}{2g''(0 | x_0) h_1^v} c_p^* \right\} \{1 + o(1)\}.$$

Furthermore, the asymptotic bias for  $p - v$  even is

$$b_v(x_0) = e_{v+1}^T S^{-1} \left[ \tilde{c}_p \frac{h_1^{p+2-v} v!}{(p+2)!} \left\{ m^{(p+2)}(x_0) + (p+2) m^{(p+1)}(x_0) \frac{\Gamma'(x_0)}{\Gamma(x_0)} \right\} - \frac{g'''(0 | x_0) v! h_2^2}{2g''(0 | x_0) h_1^v} c_p^* \right] \{1 + o(1)\}, \quad (2.5)$$

provided that  $m^{(p+2)}(\cdot)$  is continuous in a neighborhood of  $x_0$ , where  $\Gamma(x) = g''(0 | x)f(x)$ .

The proof of Theorem 2.2 is given in the Appendix. Similar to the local polynomial regression (LPR), the second term in (2.5) often creates extra bias and depends on the design density  $f(x)$ . Thus, it is preferable to use odd values of  $p - v$  in practice. Therefore, it is consistent with the selection order of  $p$  for the LPR (Fan and Gijbels, 1996).



**Theorem 2.3.** *Under the regularity conditions (A1)–(A4) in the Appendix, if the bandwidths  $h_1$  and  $h_2$  go to 0 such that  $nh_1h_2^5 \rightarrow \infty$  and  $h_1^{p+1}/h_2 \rightarrow 0$ , the estimate  $\hat{m}_v(x_0)$ , given in Theorem 2.1, has the following asymptotic distribution*

$$\frac{\hat{m}_v(x_0) - m^{(v)}(x_0) - b_v(x_0)}{\sqrt{\text{Var}\{\hat{m}_v(x_0)\}}} \xrightarrow{D} N(0, 1).$$

The proof of Theorem 2.3 is given in the Appendix. Specially, when  $p = 1$  and  $v = 0$ , the asymptotic variance of  $\hat{m}(x_0)$  is

$$\text{Var}\{\hat{m}(x_0)\} \approx \frac{g(0 | x_0)\tilde{\nu}\nu_0}{nh_1h_2^3g''(0 | x_0)^2f(x_0)} \{1 + o_p(1)\},$$

and the asymptotic bias is

$$b(x_0) \approx \frac{1}{2}m''(x_0)\mu_2h_1^2 - \frac{g'''(0 | x_0)h_2^2}{2g''(0 | x_0)}.$$

To find the global optimal bandwidth, we proposed to minimize the asymptotic weighted mean integrated squared error given by

$$\int ([\text{Bias}\{\hat{m}(x_0) | X\}]^2 + \text{Var}\{\hat{m}(x_0) | X\}) w(x)dx = \frac{K}{nh_1h_2^3} + Mh_1^4 + Nh_2^4 + 2Lh_1^2h_2^2,$$

where

$$\begin{aligned} K &= \int \frac{g(0 | x)\tilde{\nu}\nu_0}{g''(0 | x)^2f(x)} w(x)dx, & M &= \int \left[ \frac{1}{2}m''(x)\mu_2 \right]^2 w(x)dx, \\ N &= \int \left[ -\frac{g'''(0 | x)}{2g''(0 | x)} \right]^2 w(x)dx, & L &= \int \left[ \frac{1}{2}m''(x)\mu_2 \right] \left[ -\frac{g'''(0 | x)}{2g''(0 | x)} \right] w(x)dx, \end{aligned}$$

and  $w(x)$  is a weight function, such as 1 or the design density  $f(x)$ . Therefore, the

asymptotic global optimal bandwidth is

$$\hat{h}_1 = \left[ \frac{3K}{4n\delta^5(L + N\delta^2)} \right]^{1/8}, \hat{h}_2 = \delta\hat{h}_1, \quad (2.6)$$

where  $\delta^2 = (\sqrt{L^2 + 3MN} + L)/N$ .

### 3. NONPARAMETRIC VARYING COEFFICIENT MODAL REGRESSION

Next, we will introduce how to apply the idea of varying-coefficient models to modal regression. Varying coefficient models (Hastie and Tibshirani, 1993) have been successfully applied to high-dimensional nonparametric regression, generalized linear models, longitudinal and functional data analysis, and others. Please see, for example, Hoover, Rice, Wu, and Yang (1998); Wu, Chiang, and Hoover (1998); Fan and Zhang (1999, 2000); Chiang, Rice, and Wu (2001); Huang, Wu, and Zhou (2002), for more details.

Given a random sample  $\{(\mathbf{x}_i, u_i, y_i), i = 1, \dots, n\}$ , where  $y_i$  is the response variable,  $\mathbf{x}_i$  is a  $p$ -dimensional predictor (with first component equals 1), and  $u_i$  is a scalar predictor. Suppose  $f(y | \mathbf{x}_i, u_i)$  is the conditional density function of  $y$  given  $(\mathbf{x}_i, u_i)$ . The *varying coefficient modal regression* assumes

$$\text{Mode}(y | \mathbf{x}_i, u_i) = \max_y f(y | \mathbf{x}_i, u_i) = \sum_{j=1}^p g_j(u_i) x_{ij}, \quad (3.1)$$

where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$  and  $\{g_1(u), \dots, g_p(u)\}^T$  are unknown smooth functions. Note that nonparametric modal regression (2.1) is a special case of (3.1) if we take  $p = 1$  and  $x_i = 1$ . The proposed method can be easily extended to the case when  $U$  is multivariate. However, the extension to the multivariate  $U$  might be practically less useful due to the “curse of dimensionality”.

We propose to estimate the varying coefficient modal regression by *maximizing*

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^n K_{h_1}(u_i - u_0) \phi_{h_2} \left[ y_i - \sum_{j=1}^p \{b_j + c_j(u_i - u_0)\} x_{ij} \right], \quad (3.2)$$

where  $\boldsymbol{\theta} = (b_1, \dots, b_p, h_1 c_1, \dots, h_1 c_p)^T$ . For the simplicity of explanation, here we only consider the local linear approximation based on the arguments following Theorem 2.2.

We can use an algorithm similar to the EM algorithm proposed in Section 2.2 to maximize (3.2). Starting with  $k = 0$ :

*E-Step:* Update  $\pi(j \mid \boldsymbol{\theta}^{(k)})$

$$\pi(j \mid \boldsymbol{\theta}^{(k)}) = \frac{K_{h_1}(u_j - u_0) \phi_{h_2} \left[ y_j - \sum_{l=1}^p \{b_l^{(k)} + c_l^{(k)}(u_j - u_0)\} x_{jl} \right]}{\sum_{i=1}^n K_{h_1}(u_i - u_0) \phi_{h_2} \left[ y_i - \sum_{l=1}^p \{b_l^{(k)} + c_l^{(k)}(u_i - u_0)\} x_{il} \right]}, j = 1, \dots, n.$$

*M-Step:* Update  $\boldsymbol{\theta}^{(k+1)}$

$$\boldsymbol{\theta}^{(k+1)} = \arg \max_{\boldsymbol{\theta}} \sum_{j=1}^n \pi(j \mid \boldsymbol{\theta}^{(k)}) \log \phi_{h_2} \left[ y_j - \sum_{l=1}^p \{b_l^{(k)} + c_l^{(k)}(u_j - u_0)\} x_{jl} \right],$$

which has explicit solution since  $\phi(\cdot)$  is the Gaussian density.

Denote by  $f(u)$  the marginal density of  $u$ ,  $q(\epsilon \mid \mathbf{x}, u)$  the conditional density of  $\epsilon = y - \sum_{j=1}^p g_j(u) x_j$  given  $\mathbf{x}$  and  $u$ , and  $q^{(v)}(\epsilon \mid \mathbf{x}, u)$  the  $v$ -th derivative of  $q(\epsilon \mid \mathbf{x}, u)$ .

Let

$$\begin{aligned} \alpha_j(u) &= \mathbb{E}\{\mathbf{x} X_j q^{(2)}(0 \mid \mathbf{x}, u) \mid u\}, & \beta(u) &= \mathbb{E}\{\mathbf{x} q^{(3)}(0 \mid \mathbf{x}, u) \mid u\} \\ \Delta(u) &= \mathbb{E}\{\mathbf{x} \mathbf{x}^T q^{(2)}(0 \mid \mathbf{x}, u) \mid u\}, & \tilde{\Delta}(u) &= \mathbb{E}\{\mathbf{x} \mathbf{x}^T q(0 \mid \mathbf{x}, u) \mid u\}. \end{aligned}$$

Suppose  $\hat{\boldsymbol{\theta}}$  is the maximizer of (3.2), then  $\hat{\mathbf{g}}(u_0) = (\hat{b}_1, \dots, \hat{b}_p)^T$  is the estimate of  $\{g_1(u_0), \dots, g_p(u_0)\}^T$ , and  $\hat{\mathbf{g}}'(u_0) = (\hat{c}_1, \dots, \hat{c}_p)^T$  is the estimate of  $\{g'_1(u_0), \dots, g'_p(u_0)\}^T$ .

**Theorem 3.1.** *Under the regularity conditions (A5)—(A8) in the Appendix, if the bandwidths  $h_1$  and  $h_2$  go to 0 such that  $nh_1^3h_2^5 \rightarrow \infty$  and  $h_1^2/h_2 \rightarrow 0$  the asymptotic bias of  $\hat{\mathbf{g}}(u_0)$  is given by*

$$\text{Bias}\{\hat{\mathbf{g}}(u_0)\} = \frac{1}{2}\Delta^{-1}(u_0) \left\{ \mu_2 h_1^2 \sum_{j=1}^p g_j''(u_0) \alpha_j(u_0) - h_2^2 \beta(u_0) \right\} \{1 + o_p(1)\}, \quad (3.3)$$

and the asymptotic covariance is

$$\text{Cov}\{\hat{\mathbf{g}}(u_0)\} = \frac{\tilde{\nu}\nu_0}{nh_1h_2^3f(u_0)}\Delta^{-1}(u_0)\tilde{\Delta}(u_0)\Delta^{-1}(u_0) \{1 + o_p(1)\}, \quad (3.4)$$

where  $\mu_j = \int t^j K(t)dt$ ,  $\nu_j = \int t^j K^2(t)dt$ , and  $\tilde{\nu} = \int t^2 \phi^2(t)dt$ .

**Theorem 3.2.** *Under regularity conditions (A5)—(A8) in the Appendix, if the bandwidths  $h_1$  and  $h_2$  go to 0 such that  $nh_1h_2^5 \rightarrow \infty$  and  $h_1^2/h_2 \rightarrow 0$ , the estimate  $\mathbf{g}(u_0)$  has the following asymptotic distribution*

$$[\text{Cov}\{\hat{\mathbf{g}}(u_0)\}]^{-1/2}[\hat{\mathbf{g}}(u_0) - \mathbf{g}_0(u_0) - \text{Bias}\{\hat{\mathbf{g}}(u_0)\}] \xrightarrow{D} N(0, I),$$

where  $\text{Bias}\{\hat{\mathbf{g}}(u_0)\}$  is defined in (3.3) and  $\text{Cov}\{\hat{\mathbf{g}}(u_0)\}$  is defined in (3.4).

The asymptotic global optimal bandwidth can be found by minimizing the asymptotic weighted mean integrated squared error given by

$$\begin{aligned} & \int \text{E} [\{\hat{\mathbf{g}}(u) - \mathbf{g}_0(u)\}^T W \{\hat{\mathbf{g}}(u) - \mathbf{g}_0(u)\}] w(u) du \\ & \approx \int [\text{Bias}\{\hat{\mathbf{g}}(u)\}^T W \text{Bias}\{\hat{\mathbf{g}}(u)\} + \text{tr} [\text{Cov}\{\hat{\mathbf{g}}(u)\} W]] w(u) du, \end{aligned}$$

where  $W$  is a weight matrix and  $w(u)$  is a weight function, such as 1 or the design density for  $u$ . One popular choice for  $W$  is  $[\Delta^{-1}(U)\tilde{\Delta}(U)\Delta^{-1}(U)]^{-1}$ , which is proportional to the inverse of the asymptotic variance of  $\hat{\mathbf{g}}(u)$ . Based on the asymptotic results of (3.3) and (3.4), the theoretical global optimal bandwidths are

$$\hat{h}_1 = \left[ \frac{3\tilde{K}}{4n\delta^5(\tilde{L} + \tilde{N}\tilde{\delta}^2)} \right]^{1/8}, \quad \hat{h}_2 = \tilde{\delta}\hat{h}_1,$$

where

$$\begin{aligned} \tilde{K} &= p\tilde{\nu}\nu_0 \int f(u)^{-1}w(u)du, \quad \tilde{L} = -\mu_2 \int \sum_{j=1}^p \alpha_j(u)\tilde{\Delta}(u)^{-1}\beta(u)w(u)du \\ \tilde{N} &= \int \beta(u)'\tilde{\Delta}(u)\beta(u)w(u)du, \quad \tilde{\delta}^2 = (\sqrt{\tilde{L}^2 + 3\tilde{M}\tilde{N}} + \tilde{L})/\tilde{N}, \\ \tilde{M} &= \mu_2^2 \int \left\{ \sum_{j=1}^p g_j''(u)\alpha_j'(u) \right\} \tilde{\Delta}^{-1}(u) \left\{ \sum_{j=1}^p g_j''(u)\alpha_j'(u) \right\} w(u)du. \end{aligned}$$

We will also investigate how to practically choose the bandwidth based on the above theoretical results in the simulation study.

#### 4. SIMULATION STUDY AND APPLICATION

In this section, we will use a Monte Carlo simulation study and a real data application to assess the performance of the proposed nonparametric modal regression.

To use the proposed two nonparametric modal regression models, we need to select the bandwidths first. Note that the asymptotic global optimal bandwidth formula (2.6) contains the unknown quantities  $m''(x)$  and  $g^{(\nu)}(0|x)$ ,  $\nu = 0, 2, 3$ , the  $\nu$ -th derivative of conditional density of  $\epsilon$  given  $x$ , and therefore, is not ready to use. One possible practical way is to apply the plug-in method by replacing the unknown quantities with some estimates. We propose to estimate  $m(x)$  by a polynomial function of order three, i.e.,  $m(x) \approx \tilde{\mathbf{x}}^T \boldsymbol{\alpha}$ , where  $\tilde{\mathbf{x}} = (1, x, x^2, x^3)^T$  and  $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^T$ .

We can then estimate  $\epsilon_i$  by  $\hat{\epsilon}_i = y_i - \tilde{\mathbf{x}}_i^T \hat{\boldsymbol{\alpha}}$  and  $m''(x)$  by  $\hat{m}''(x) = 2\hat{\alpha}_2 + 6\hat{\alpha}_3 x$ , where  $\hat{\boldsymbol{\alpha}}$  is the modal linear regression estimator (Yao and Li, 2013). In our simulation,  $\epsilon$  and  $x$  are independent. Therefore,  $\hat{\epsilon}_i - \hat{m}(x_i)$  has approximate density  $g(\cdot)$ , and  $g^{(\nu)}(0|x)$  can be estimated by

$$\hat{g}^{(\nu)}(0|x) = \frac{1}{h^{\nu+1}} \sum_{i=1}^n K^{(\nu)} \left\{ \frac{\hat{\epsilon}_i - \hat{m}(x_i)}{h} \right\}, \nu = 0, 2, 3.$$

If  $w(x)$  in (2.6) is equal to the design density  $f(x)$ , then  $K$ ,  $M$ ,  $N$ , and  $L$  can be estimated by their empirical version:

$$\begin{aligned} K &= \frac{1}{n} \sum_{i=1}^n \frac{\hat{g}(0 | x_i) \tilde{\nu} \nu_0}{\hat{g}''(0 | x_i)^2 f(x_i)}, M = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{2} \hat{m}''(x_i) \mu_2 \right\}^2, \\ N &= \frac{1}{n} \sum_{i=1}^n \left\{ -\frac{\hat{g}'''(0 | x_i)}{2 \hat{g}''(0 | x_i)} \right\}^2, L = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{2} \hat{m}''(x_i) \mu_2 \right\} \left\{ -\frac{\hat{g}'''(0 | x_i)}{2 \hat{g}''(0 | x_i)} \right\}. \end{aligned}$$

#### 4.1. Monte Carlo simulations

**Example 1:** Generate i.i.d. sample  $\{(x_i, y_i), i = 1, \dots, n\}$  from

$$Y = 2 \sin(\pi X) + \sigma(X) \epsilon,$$

with  $X \sim U(0, 1)$ , and  $\sigma(X) = 1 + 2X$ . The error is distributed as  $\epsilon \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ , such that the mean is 0, the mode is approximately 1 and the median is around 0.67. The proposed modal regression function is  $\text{Mode}(Y|X) = 2 \sin(\pi X) + 1 + 2X$ , the median regression is  $\text{Median}(Y|X) = 2 \sin(\pi X) + 0.67 + 1.34X$ , and the traditional mean regression is  $E(Y | X) = 2 \sin(\pi X)$ . We consider the following four regression estimates: 1) local linear estimate (LL); 2) local M-estimate (LM); 3) Local median regression estimate (LMD); 4) the proposed local linear modal regression (LLMR). The sample sizes  $n = 200, 400$  and  $800$  are conducted over 500

repetitions.

Note that the above four regression estimates are targeting different regression functions. That is, LL and LM are targeting the mean regression function, LMD is targeting the median regression function, and LLMR is targeting the modal regression function. In order to compare the performance of different estimates, we will look at their prediction performance based on the coverage probabilities when doing prediction based on the same length of small intervals centered around each estimate. For the error distribution assumed above, the standard deviation is close to  $\sigma = 2$ , therefore, the lengths of intervals considered are  $0.1\sigma$ ,  $0.2\sigma$ , and  $0.5\sigma$ . The coverage probabilities are approximated by doing prediction for the 1,000 equally spaced grid points from 0.1 to 0.9, with 500 repetitions.

Table 1 contains the average and standard deviation of the estimated coverage probabilities when doing prediction based on the same length of intervals centered around each estimate. From Table 1 we can see that LLMR provides the highest coverage probability among the methods considered. In addition, LMD also provides better prediction performance than the mean regression estimates LL and LM, partly due to the skewness of the error distribution.

**Example 2:** In this example, we demonstrate the performance of the proposed nonparametric varying coefficient modal regression by the following two models:

**Model 1:**  $y = g_0(u) + g_1(u)x_1 + g_2(u)x_2 + \sigma(u)\epsilon$ , where  $g_0(u) = \exp(2u - 1)$ ,  $g_1(u) = 8u(1 - u)$ , and  $g_2(u) = 2 \sin^2(2\pi u)$ .

**Model 2:**  $y = g_0(u) + g_1(u)x_1 + g_2(u)x_2 + \sigma(u)\epsilon$ , where  $g_0(u) = \sin(2\pi u)$ ,  $g_1(u) = (2u - 1)^2 + 0.5$ , and  $g_2(u) = \exp(2u - 1) - 1$ .

In both models,  $x_1$  and  $x_2$  follow a standard normal distribution with correlation coefficient  $1/\sqrt{2}$ . The index variable  $u$  is uniformly distributed on  $[0, 1]$ , and is independent of  $(x_1, x_2)$ . Similar to the previous example, we consider  $\epsilon \sim$

Table 1: Average (Std) of percentage of coverage with  $\sigma = 2$ .

Width	Method	n=200	n=400	n=800
0.1 $\sigma$	LL	0.030(0.007)	0.029(0.005)	0.028(0.004)
	LM	0.047(0.011)	0.047(0.008)	0.047(0.007)
	LMD	0.073(0.009)	0.075(0.007)	0.077(0.005)
	LLMR	0.081(0.014)	0.086(0.011)	0.090(0.009)
0.2 $\sigma$	LL	0.062(0.014)	0.059(0.011)	0.057(0.008)
	LM	0.095(0.021)	0.095(0.017)	0.095(0.013)
	LMD	0.145(0.017)	0.149(0.012)	0.153(0.009)
	LLMR	0.160(0.026)	0.169(0.021)	0.176(0.017)
0.5 $\sigma$	LL	0.173(0.033)	0.169(0.026)	0.166(0.020)
	LM	0.248(0.044)	0.252(0.034)	0.254(0.027)
	LMD	0.340(0.031)	0.350(0.020)	0.356(0.014)
	LLMR	0.366(0.046)	0.382(0.036)	0.393(0.027)

$0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ , and  $\sigma(u) = 1 + 2u$ .

We conduct simulations with sample sizes  $n = 200, 400$ , and  $800$ , respectively, with 200 data replications. The bandwidths for LL, LM, and LMD are chosen by cross-validation.

To compare the coverage probabilities of all methods, we take 30 equally spaced points from 0.1 to 0.9 for  $x_1$ ,  $x_2$ , and  $u$ , and do prediction for all of the 9,000 grid points. Tables 2 and 3 contain the estimated coverage probabilities for Model 1 and Model 2, respectively, based on the same length of small intervals centered around each estimate. From Tables 2 and 3, we can see that LLMR provides higher coverage probabilities than all the other three methods, which becomes even more obvious when the sample size increases. In addition, LMD and LM also provide higher coverage probabilities than LL.

#### 4.2. Health Care Expenditure data.



Table 2: Average (Std) of percentage of coverage of Model 1, with  $\sigma = 2$ .

Width	Method	n=200	n=400	n=800
0.1 $\sigma$	LL	0.031(0.008)	0.028(0.005)	0.027(0.003)
	LM	0.041(0.010)	0.040(0.007)	0.036(0.005)
	LMD	0.047(0.010)	0.043(0.007)	0.041(0.005)
	LLMR	0.067(0.013)	0.076(0.012)	0.080(0.013)
0.2 $\sigma$	LL	0.062(0.015)	0.058(0.010)	0.056(0.007)
	LM	0.084(0.020)	0.081(0.014)	0.075(0.010)
	LMD	0.094(0.020)	0.088(0.014)	0.084(0.010)
	LLMR	0.133(0.026)	0.150(0.022)	0.158(0.024)
0.5 $\sigma$	LL	0.171(0.037)	0.164(0.025)	0.163(0.018)
	LM	0.222(0.044)	0.220(0.031)	0.208(0.024)
	LMD	0.243(0.041)	0.233(0.030)	0.226(0.022)
	LLMR	0.320(0.049)	0.353(0.038)	0.366(0.043)

We illustrate the proposed methodology by an analysis of the health care expenditure data (Cohen, 2003; Natarajan et al., 2008). The data set comes from the Medical Expenditure Panel Survey (MEPS) for the year 2002, which was conducted by the United States National Center for Health Statistics, Centers for Disease Control and Prevention. The survey was designed to produce national and regional estimates of the health care use, expenditures, sources of payment and insurance coverage of the US civilian non-institutionalized population. Medical cost data are typically highly skewed to the right, in that a small percentage of subjects sustain extremely high costs compared to other subjects.

We randomly select 500 patients within one “primary sampling units” (PSUs) as our data example. The outcome of interest is “total health care expenditures in the year 2002”. The covariates of interest are race (1 if white; 0 if otherwise), smoke (1 if a current smoker; 0 if otherwise), pov (1 if above the poverty line; 0 if at or below

Table 3: Average (Std) of percentage of coverage of Model 2, with  $\sigma = 2$ .

Width	Method	n=200	n=400	n=800
$0.1\sigma$	LL	0.033(0.008)	0.030(0.006)	0.029(0.004)
	LM	0.047(0.011)	0.042(0.008)	0.040(0.006)
	LMD	0.052(0.010)	0.050(0.008)	0.049(0.005)
	LLMR	0.068(0.014)	0.077(0.012)	0.084(0.012)
$0.2\sigma$	LL	0.066(0.017)	0.062(0.012)	0.059(0.007)
	LM	0.094(0.021)	0.085(0.017)	0.082(0.011)
	LMD	0.106(0.019)	0.102(0.016)	0.099(0.010)
	LLMR	0.135(0.027)	0.152(0.024)	0.166(0.022)
$0.5\sigma$	LL	0.184(0.039)	0.177(0.030)	0.171(0.019)
	LM	0.248(0.045)	0.230(0.036)	0.225(0.025)
	LMD	0.271(0.038)	0.266(0.032)	0.261(0.021)
	LLMR	0.323(0.052)	0.355(0.042)	0.378(0.038)

the poverty line), insur (1 if the patient has health insurance; 0 if otherwise), phealth (1 if good perceived health status; 0 if otherwise) and meds (1 if the patient needs prescription medication; 0 if otherwise), and we take  $u=\text{age}$ . We fit the data by LL, LM, LMD, and LLMR.

With 10%, 30%, 50%, and 90% as the levels of confidence, Table 4 reports the average widths and percentage of coverage of the prediction intervals. The confidence interval of LLMR is constructed based on the similar method suggested by Yao and Li (2014), which could make use of the skewness of the error distribution assumed by LLMR. The coverage probability is measured by leave-one-out cross validation. From Table 4, we can see that the actual coverage rates are very close to the nominal confidence levels for all methods. The average widths of LMD and LLMR are shorter compared to LL and LM, and LLMR is superior for higher confidence levels.

To evaluate the prediction performance of the methods, we apply  $d$ -fold cross-

validation and Monte-Carlo cross-validation (MCCV) to the data, and the median and standard deviation of the median of squared prediction errors (MSPE) are reported in Table 5. The medians of LLMR are much smaller than the other three methods, indicating that LLMR provides the best point prediction followed by LMD and LM.

Table 4: Average widths (percentage of coverage) of the prediction intervals for health care expenditure data.

Method	10%	30%	50%	90%
LL	0.095(0.098)	0.255(0.290)	0.438(0.470)	1.904(0.886)
LM	0.080(0.104)	0.215(0.272)	0.394(0.486)	1.897(0.888)
LMD	0.053(0.094)	0.163(0.252)	0.381(0.438)	1.758(0.876)
LLMR	0.057(0.096)	0.172(0.276)	0.334(0.470)	1.719(0.902)

Table 5: Median (Std) of MSPE for health care expenditure data.

Method	5-fold CV	10-fold CV	MCCV d=50	MCCV d=100
LL	0.105(0.020)	0.102(0.029)	0.106(0.033)	0.108(0.023)
LM	0.079(0.018)	0.072(0.027)	0.079(0.023)	0.079(0.015)
LMD	0.042(0.010)	0.042(0.018)	0.041(0.020)	0.043(0.013)
LLMR	0.021(0.009)	0.021(0.010)	0.020(0.020)	0.025(0.016)

## 5. CONCLUDING REMARKS

In this article, we proposed a nonparametric modal regression and a nonparametric varying coefficient modal regression. Compared to traditional mean regression models, the new nonparametric modal regression models are more robust and have better prediction performance. We demonstrated such superior performance through a simulation study and a health care expenditure data.

Choosing the bandwidths has long been a difficult problem for nonparametric and semiparametric models. In this paper, we propose to use the plug-in method

to choose the bandwidths based on the found asymptotic optimal bandwidths. One might also use a sequence of bandwidths as suggested by Kemp and Santos Silva (2012) to reveal some more interesting features of modal regression. In addition, it is also interesting to know how to adapt the traditional cross validation technique to choose the bandwidth for nonparametric modal regression.

The development of modal regression is still in its initial stage. We believe that modal regression could be a good alternative to the mean regression and median regression and there are still much work to be done in the future. Much of the development for mean regression and median regression could have similar development for modal regression.

## APPENDIX

The conditions used by the theorems are listed below. They are not the weakest possible conditions, but they are imposed to facilitate the proofs.

### Technical Conditions:

- (A1) The  $m(x)$  has continuous  $(p + 1)^{th}$  derivative at the point  $x_0$ .
- (A2)  $g'(0 | x) = 0, g''(0 | x) < 0, g^{(v)}(t | x)$  is bounded in a neighbor of  $x_0$  and has continuous first derivative at the point  $x_0$  as a function of  $x$ , for  $v = 0, \dots, 4$ .
- (A3) The  $f(x)$  is bounded and has continuous first derivative at the point  $x_0$  and  $f(x_0) > 0$ .
- (A4)  $K(\cdot)$  is a symmetric (about 0) probability density with compact support  $[-1, 1]$ .
- (A5)  $g_j(x)$  has continuous  $2^{nd}$  derivative at the point  $x_0, j = 1, \dots, p$ .
- (A6)  $q'(0 | \mathbf{x}, u) = 0, q''(0 | \mathbf{x}, u) < 0, q^{(v)}(t | \mathbf{x}, u)$  is bounded in a neighbor of  $(\mathbf{x}_0, u_0)$  and has continuous first derivative at the point  $(\mathbf{x}_0, u_0)$  as a function of  $(\mathbf{x}, u)$ , for  $v = 0, \dots, 4$ .

(A7) The  $f(u)$  is bounded and has continuous first derivative at the point  $u_0$  and  $f(u_0) > 0$ .

Denote  $X_i^* = \{1, (X_i - x_0)/h_1, \dots, (X_i - x_0)^p/h_1^p\}^T$ ,  $H = \text{diag}\{1, h_1, \dots, h_1^p\}$ ,  $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ ,  $\boldsymbol{\theta}^* = H\boldsymbol{\theta}$ ,  $R(X_i) = m(X_i) - \sum_{j=0}^p \beta_j (X_i - x_0)^j$ , and  $K_i = K_{h_1}(x_i - x_0)$ , where  $\beta_j = m^{(j)}(x_0)/j!$ ,  $j = 0, 1, \dots, p$ .

*Proof of Theorem 2.1.* Denote  $\alpha_n = (nh_1h_2^3)^{-1/2} + h_1^{p+1} + h_2^2$ . It is sufficient to show that for any given  $\eta > 0$ , there exists a large constant  $c$  such that

$$P\{\sup_{\|\mu\|=c} \ell(\boldsymbol{\theta}^* + \alpha_n \mu) < \ell(\boldsymbol{\theta}^*)\} \geq 1 - \eta, \quad (\text{A.1})$$

where  $\ell(\boldsymbol{\theta})$  is defined in (2.3).

By using Taylor expansion, it follows that

$$\begin{aligned} \ell(\boldsymbol{\theta}^* + \alpha_n \mu) - \ell(\boldsymbol{\theta}^*) &= \frac{1}{n} \sum_{i=1}^n K_i \left\{ -\phi'_{h_2}(\epsilon_i + R(X_i)) \alpha_n \mu^T X_i^* + \frac{1}{2} \phi''_{h_2}(\epsilon_i + R(X_i)) \alpha_n^2 (\mu^T X_i^*)^2 \right. \\ &\quad \left. - \frac{1}{6} \phi'''_{h_2}(z_i) \alpha_n^3 (\mu^T X_i^*)^3 \right\} \\ &\triangleq I_1 + I_2 + I_3, \end{aligned}$$

where  $z_i$  is between  $\epsilon_i + R(X_i)$  and  $\epsilon_i + R(X_i) + \alpha_n \mu^T X_i^*$ . Note that

$$\phi'_{h_2}(t) = -\frac{t}{h_2^3} \phi\left(\frac{t}{h_2}\right), \phi''_{h_2}(t) = \frac{1}{h_2^3} \left(\frac{t^2}{h_2^2} - 1\right) \phi\left(\frac{t}{h_2}\right), \text{ and } \phi'''_{h_2}(t) = \frac{1}{h_2^4} \left\{ \frac{3t}{h_2} - \left(\frac{t}{h_2}\right)^3 \right\} \phi\left(\frac{t}{h_2}\right).$$

If  $a_n(x) = o_p(h_2)$ , and  $g^{(v)}(t | x)$  is bounded in a neighbor of  $x_0$ , we have

$$\begin{aligned} E\{\phi'_{h_2}(\epsilon + a_n(x)) | X = x\} &= -h_2^{-1} \int t\phi(t)g(th - a_n(x)|x)dt \\ &= -\left\{\frac{g'''(0|x)}{2}h_2^2 + g''(0|x)a_n(x)\right\}\{1 + o_p(1)\}. \end{aligned}$$

If  $h_1^{p+1}/h_2 \rightarrow 0$ , by directly calculating the mean and variance, we obtain

$$\begin{aligned} E(I_1) &= \alpha_n \mu^T \left\{ \frac{g'''(0|x_0)}{2} f(x_0) \check{c}_p h_2^2 - g''(0|x_0) c_p f(x_0) \frac{m^{(p+1)}(x_0)}{(p+1)!} h_1^{p+1} \right\} \\ &= O\{\alpha_n c (h_2^2 + h_1^{p+1})\}, \\ \text{Var}(I_1) &= n^{-1} \alpha_n^2 \mu^T \{g(0|x_0) f(x_0) \nu_0 S^* h_2^{-3} h_1^{-1}\} \mu \\ &= O(\alpha_n^2 (nh_1 h_2^3)^{-1} c^2), \end{aligned} \tag{A.2}$$

where  $\check{c}_p = (\mu_0, \mu_1, \dots, \mu_p)^T$ . Hence  $I_1 = O\{\alpha_n c (h_2^2 + h_1^{p+1})\} + \alpha_n c O_p((nh_1^{-1} h_2^{-3})^{-1/2}) = O_p(c\alpha_n^2)$ . If  $nh_2^5 h_1 \rightarrow \infty$ , similar to (A.2), we can prove

$$\begin{aligned} I_2 &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{2} K_i \phi''_{h_2}(\epsilon_i + R(X_i)) \alpha_n^2 \mu^T X_i^* X_i^{*T} \mu \right\} = \alpha_n^2 g''(0|x_0) f(x_0) \mu^T S \mu (1 + o_p(1)), \\ I_3 &= \frac{1}{n} \sum_{i=1}^n \left\{ -\frac{1}{6} K_i \phi'''_{h_2}(z_i) \alpha_n^3 (\mu^T X_i^*)^3 \right\} = o_p(\alpha_n^2). \end{aligned} \tag{A.3}$$

Noticing that  $S$  is a positive matrix,  $\|\mu\| = c$ , and  $g''(0|x_0) < 0$ , we can choose  $c$  large enough such that  $I_2$  dominates both  $I_1$  and  $I_3$  with probability at least  $1 - \eta$ . Thus (A.1) holds. Therefore, with probability approaching 1 (wpa1), there exists a local maximizer  $\hat{\theta}^*$  such that  $\|\hat{\theta}^* - \theta^*\| \leq \alpha_n c$ . Based on the definition of  $\theta^*$ , we can get, wpa1,  $|h_1^v \{\hat{m}_v(x_0) - m^{(v)}(x_0)\}| = O_p\{(nh_1 h_2^3)^{-1/2} + h_1^{p+1} + h_2^2\}$ .  $\square$

Define

$$W_n = \sum_{i=1}^n X_i^* K_i \phi'_{h_2}(\epsilon_i). \tag{A.4}$$

We have the following asymptotic representation.

**Lemma A.1.** *Under conditions (A1)—(A4), it follows that*

$$\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^* = h_1^{p+1} \frac{m^{(p+1)}(x_0)}{(p+1)!} S^{-1} c_p (1 + o_p(1)) + \frac{S^{-1} W_n}{n F(x_0, h_2) f(x_0)} (1 + o_p(1)). \quad (\text{A.5})$$

**Proof.** Let  $\hat{\gamma}_i = R(X_i) - \sum_{j=0}^p (\hat{\beta}_j - \beta_j)(X_i - x_0)^j = R(X_i) - (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*)^T X^*$ , then  $Y_i - \sum_{j=0}^p \hat{\beta}_j (X_i - x_0)^j = \epsilon_i + \hat{\gamma}_i$ . The solution  $\hat{\boldsymbol{\theta}}^*$  satisfies the equation

$$\sum_{i=1}^n X_i^* K_i \phi'_{h_2}(\epsilon_i + \hat{\gamma}_i) = \sum_{i=1}^n X_i^* K_i \left\{ \phi'_{h_2}(\epsilon_i) + \phi''_{h_2}(\epsilon_i) \hat{\gamma}_i + \frac{1}{2} \phi'''_{h_2}(\epsilon^*) \hat{\gamma}_i^2 \right\} = 0, \quad (\text{A.6})$$

where  $\epsilon^*$  is between  $\epsilon_i$  and  $\epsilon_i + \hat{\gamma}_i$ . Note that the second term on the left hand side of (A.6) is

$$\sum_{i=1}^n K_i \phi''_{h_2}(\epsilon_i) R(X_i) X_i^* - \sum_{i=1}^n K_i \phi''_{h_2}(\epsilon_i) X_i^* X_i^{*'} (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*) \triangleq J_1 + J_2. \quad (\text{A.7})$$

From the proof of (A.2), we have

$$J_1 = n h_1^{p+1} g''(0 | x_0) f(x_0) c_p \frac{m^{(p+1)}(x_0)}{(p+1)!} + o_p(n h_1^{p+1}),$$

and

$$J_2 = -n g''(0 | x_0) f(x_0) S (1 + o_p(1)) (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*).$$

From Theorem 2.1, we know  $\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\| = O_p\{h_1^{p+1} + h_2^2 + (n h_1 h_2^3)^{-1/2}\}$ , hence

$$\begin{aligned} \sup_{i: |X_i - x_0|/h \leq 1} |\hat{\gamma}_i| &\leq \sup_{i: |X_i - x_0|/h \leq 1} |R(X_i)| + (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*)^T X^* \\ &= O_p(h_1^{p+1} + \|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|) = O_p(\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|). \end{aligned} \quad (\text{A.8})$$

$(m^{(p+1)}(x))$  is bounded) Also, we have

$$\begin{aligned} \mathbb{E} \{K_i(X_i - x_0)^j / h_1^j\} &= \int \frac{1}{h_1} K\left(\frac{x - x_0}{h_1}\right) \left(\frac{x - x_0}{h}\right)^j f(x) dx \\ &= \mu_j f(x_0) + o(1). \end{aligned}$$

Based on (A.3), (A.8), and  $\|\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\| = O_p(\alpha_n)$ ,

$$\frac{1}{n} \left\{ \sum_{i=1}^n K_i \hat{\gamma}_i^2 \phi_{h_2}'''(\epsilon^*)(X_i - x_0)^j / h_1^j \right\} = \frac{1}{n} \left\{ \sum_{i=1}^n K_i \alpha_n^2 \phi_{h_2}'''(\epsilon^*)(X_i - x_0)^j / h_1^j \right\} = o_p(\|\alpha_n\|).$$

Hence for the third term on the left-hand side of (A.6),

$$\sum_{i=1}^n K_i \hat{\gamma}_i^2 X_i^* \phi_{h_2}'''(\epsilon^*) = o_p(n\alpha_n) = o_p(J_2).$$

Then, it follows from (A.4) and (A.6) that

$$\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^* = h_1^{p+1} \frac{m^{(p+1)}(x_0)}{(p+1)!} S^{-1} c_p (1 + o_p(1)) + \frac{S^{-1} W_n}{n g''(0 | x_0) f(x_0)} (1 + o_p(1)).$$

□

*Proof of Theorem 2.2.* Based on (A.4) and the condition (A6), we can easily get

$$\mathbb{E} \left( \frac{1}{n} W_n \right) = \mathbb{E} \{ X_i^* K_i \phi_{h_2}'(\epsilon_i) \} = - \frac{g'''(0 | x_0)}{2} f(x_0) c_p^* h_2^2 + o_p(h_2^2).$$

Similar to the proof of (A.2), we have

$$\mathbb{E} \{ K_i^2 \phi_{h_2}'(\epsilon_i)^2 (X_i - x_0)^j / h_1^j \} = (h_1 h_2^3)^{-1} \nu_j \tilde{\nu} g(0 | x_0) f(x_0) \{1 + o(1)\}.$$



where  $\tilde{\nu} = \int \phi^2(t)t^2 dt$ . So

$$\text{Cov}(W_n/n) = (nh_1 h_2^3)^{-1} \tilde{\nu} g(0 | x_0) f(x_0) S^* (1 + o(1)). \quad (\text{A.9})$$

Based on the result (A.5), the asymptotic bias  $b_v(x_0)$  and variance of  $\hat{m}_v(x_0)$  are naturally given by

$$b_v(x_0) = e_{v+1}^T S^{-1} \left\{ h_1^{p+1-v} \frac{v!}{(p+1)!} m^{(p+1)}(x_0) c_p - \frac{g'''(0 | x_0) v! h_2^2}{2g''(0 | x_0) h_1^v} c_p^* \right\} (1 + o(1))$$

and

$$\text{Var}\{\hat{m}_v(x_0)\} = \frac{v!^2 g(0 | x_0) \tilde{\nu}}{nh_2^3 h_1^{1+2v} f(x_0) g''(0 | x_0)^2} e_{v+1}^T S^{-1} S^* S^{-1} e_{v+1} (1 + o(1)).$$

Noting that  $\mu_j = 0$  for odd  $j$ , by some simple calculation, we can know the  $(v+1)^{\text{th}}$  element of  $S^{-1} c_p$  is zero for  $p-v$  even. So we need higher order expansion of asymptotic bias for  $p-v$  even. Following the similar arguments of Theorem 2.1, if  $nh_1^3 h_2^5 \rightarrow \infty$  (make the root of variance order less than bias order), we can prove

$$\begin{aligned} J_1 &= nh_1^{p+1} \left[ \Gamma(x_0) c_p \frac{m^{(p+1)}(x_0)}{(p+1)!} + h_1 \tilde{c}_p \left\{ \Gamma'(x_0) \frac{m^{(p+1)}(x_0)}{(p+1)!} + \Gamma(x_0) \frac{m^{(p+2)}(x_0)}{(p+2)!} \right\} \right] \{1 + o_p(1)\}, \\ J_2 &= -n \left\{ \Gamma(x_0) S + h_1 \tilde{S} \Gamma'(x_0) \right\} \{1 + o_p(1)\} (\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*), \end{aligned}$$

where  $J_1$  and  $J_2$  are defined in (A.7) and  $\Gamma(x) = g''(0 | x) f(x)$ .

Then, it follows from (A.6) that

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = h_1^{p+1} \left\{ \frac{m^{(p+1)}(x_0)}{(p+1)!} S^{-1} c_p + h_1 b^*(x_0) \right\} (1 + o_p(1)) + \frac{S^{-1} W_n}{n \Gamma(x_0)} (1 + o_p(1)),$$

where

$$b^*(x_0) = \Gamma^{-1}(x_0) S^{-1} \tilde{c}_p \left\{ \Gamma'(x_0) \frac{m^{(p+1)}(x_0)}{(p+1)!} + \Gamma(x_0) \frac{m^{(p+2)}(x_0)}{(p+2)!} \right\} \\ - \Gamma^{-1}(x_0) \Gamma'(x_0) \frac{m^{(p+1)}(x_0)}{(p+1)!} S^{-1} \tilde{S} S^{-1} c_p.$$

For  $p - v$  even, since the  $(v+1)^{th}$  element of  $S^{-1}c_p$  and  $S^{-1}\tilde{S}S^{-1}c_p$  are zeros, the asymptotic bias  $b_v(x_0)$  of  $\hat{m}_v(x_0)$  are naturally given by

$$b_v(x_0) = e_{v+1}^T S^{-1} \left[ \tilde{c}_p \frac{h_1^{p+2-v} v!}{(p+2)!} \left\{ m^{(p+2)}(x_0) + (p+2) m^{(p+1)}(x_0) \frac{\Gamma'(x_0)}{\Gamma(x_0)} \right\} \right. \\ \left. - \frac{g'''(0 | x_0) v! h_2^2}{2g''(0 | x_0) h_1^v} c_p^* \right] \{1 + o(1)\}.$$

□

*Proof of Theorem 2.3.*

It is sufficient to show that

$$W_n^* \equiv \sqrt{h_1 h_2^3 / n} W_n \xrightarrow{D} N(0, D), \quad (\text{A.10})$$

where  $D = \tilde{v} g(0 | x_0) f(x_0) S^*$ , because using Slutsky's theorem, it follows from (A.5), (A.10), and Theorem 2.2 that

$$\frac{\hat{m}_v(x_0) - m^{(v)}(x_0) - b_v(x_0)}{\sqrt{\text{Var}\{\hat{m}_v(x_0)\}}} \xrightarrow{D} N(0, 1).$$

Next we show (A.10). For any unit vector  $d \in \mathbb{R}^{p+1}$ , we prove

$$\{d^T \text{Cov}(W_n^*) d\}^{-\frac{1}{2}} \{d^T W_n^* - d^T E(W_n^*)\} \xrightarrow{D} N(0, 1).$$

Let

$$\xi_i = \sqrt{h_1 h_2^3 / n} K_i \phi'_{h_2}(\epsilon_i) d^T X_i^*.$$

Then  $d^T W_n^* = \sum_{i=1}^n \xi_i$ . We check the Lyapunov's condition. Based on (A.9), we can get  $\text{Cov}(W_n^*) = \tilde{\nu} g(0 | x_0) f(x_0) S^* (1 + o(1))$  and  $\text{Var}(d^T W_n^* d) = d^T \text{Cov}(W_n^*) d = \tilde{\nu} g(0 | x_0) f(x_0) d^T S^* d (1 + o(1))$ . So we only need to prove  $n E |\xi_1|^3 \rightarrow 0$ . Noticing that  $(d^T X_i)^2 \leq \|d\|^2 \|X_i\|^2$ ,  $\phi'(\cdot)$  is bounded, and  $K(\cdot)$  has compact support,

$$n E |\xi|^3 \leq O(n n^{-3/2} h_1^{3/2} h_2^{9/2}) \sum_{j=0}^p E \left| K_1^3 \phi'_{h_2}(\epsilon_1)^3 \left( \frac{X_1 - x_0}{h_1} \right)^{3j} \right| \rightarrow 0.$$

So the asymptotic normality for  $W_n^*$  holds with covariance matrix  $\tilde{\nu} g(0 | x_0) f(x_0) S^*$ .  $\square$

*Proof of Theorem 3.1.* The proof is similar to Theorem 2.2 and Theorem 2.3. Here, we provide a sketch of the proof. Let  $\theta_0$  be the true value of  $\theta$ . Note that when  $u_i$  is close to  $u_0$ , we have

$$\begin{aligned} y_i - \tilde{x}_i^T \theta_0 &= \epsilon_i + m(\mathbf{x}_i, u_i) - \sum_{j=1}^p \{g_j(u_0) + g'_j(u_0)(u_i - u_0)\} x_{ij} \\ &\triangleq \epsilon_i + R(\mathbf{x}_i, u_i) = \epsilon_i + \frac{1}{2} \sum_{j=1}^p g''_j(u_0)(u_i - u_0)^2 x_{ij} (1 + o_p(1)). \end{aligned}$$

For simplicity of notations, we denote  $R(\mathbf{x}_i, u_i)$  by  $R_i$ , and  $q(\epsilon | \mathbf{x}_i, u_i)$  by  $q_i(\epsilon)$ . Then the objective function is

$$\ell(\theta_0) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(u_i - u_0) \phi_{h_2}(y_i - \tilde{\mathbf{x}}_i^T \theta_0) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(u_i - u_0) \phi_{h_2}(\epsilon_i + R_i).$$

Note that

$$\begin{aligned} \mathbb{E} [\phi''_{h_2}(\epsilon_i + R_i) \mid \mathbf{x}_i, u_i] &= q''_i(0)(1 + o_p(1)), \\ \mathbb{E} [\phi'_{h_2}(\epsilon_i + R_i) \mid \mathbf{x}_i, u_i] &= - [q'''_i(0)h_2^2/2 - q''_i(0)R_i] (1 + o_p(1)), \end{aligned}$$

and

$$\ell'(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(u_i - u_0) \phi'_{h_2}(\epsilon_i + R_i) \tilde{\mathbf{x}}_i.$$

Let  $\alpha_j(u_i) = \mathbb{E}\{\mathbf{x}_i x_{ij} q''_i(0) \mid u_i\}$  and  $\beta(u_i) = \mathbb{E}\{\mathbf{x}_i q'''_i(0) \mid u_i\}$ , then we have

$$\mathbb{E}\{\ell'(\boldsymbol{\theta}_0)\} = \left\{ -\frac{h_2^2}{2} f(u_0) \begin{pmatrix} 1 \\ \mu_1 \end{pmatrix} \otimes \beta(u_0) + \frac{h_1^2}{2} \sum_{j=1}^p g''_j(u_0) f(u_0) \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \otimes \alpha_j(u_0) \right\} (1 + o_p(1)).$$

Note that

$$\ell''(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^n K_{h_1}(u_i - u_0) \phi''_{h_2}(\epsilon_i + R_i) \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T,$$

then we have

$$\mathbb{E}\{\ell''(\boldsymbol{\theta}_0)\} = f(u_0) \begin{pmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \otimes \Delta(u_0) (1 + o_p(1)).$$

In addition, since

$$\mathbb{E}\{\phi'_{h_2}(\epsilon_i + R_i)^2 \mid u_i, \mathbf{x}_i\} = h_2^{-3} \int t^2 \phi^2(t) q_i(th - R_i) dt = h_2^{-3} q_i(0) \tilde{\nu} (1 + o_p(1)),$$

then,

$$\begin{aligned}\text{Var}\{\ell'(\boldsymbol{\theta}_0)\} &= \frac{\tilde{\nu}}{nh_2^3} \mathbb{E}[K_{h_1}^2(u_i - u_0)q_i(0)\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^T](1 + o_p(1)) \\ &= \frac{\tilde{\nu}}{nh_2^3h_1} f(u_0) \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \tilde{\Delta}(u_0)(1 + o_p(1)).\end{aligned}\quad (\text{A.11})$$

Therefore,

$$\text{Bias}\left\{\begin{pmatrix} \hat{\mathbf{g}}(u_0) \\ \hat{\mathbf{g}}'(u_0) \end{pmatrix}\right\} = \frac{1}{2}\Delta^{-1}(u_0) \otimes \begin{pmatrix} \mu_2 h_1^2 \sum_{j=1}^p g_j''(u_0)\alpha_j(u_0) - h_2^2 \beta(u_0) \\ \mu_3 h_1^2 \sum_{j=1}^p g_j''(u_0)\alpha_j(u_0) - \mu_1 h_2^2 \beta(u_0) \end{pmatrix} \{1 + o_p(1)\},$$

and

$$\text{Cov}\left\{\begin{pmatrix} \hat{\mathbf{g}}(u_0) \\ \hat{\mathbf{g}}'(u_0) \end{pmatrix}\right\} = \frac{\tilde{\nu}}{nh_1 h_2^3 f(u_0)} \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \Delta^{-1}(u_0) \tilde{\Delta}(u_0) \Delta^{-1}(u_0) (1 + o_p(1)).$$

Specifically,

$$\text{Bias}\{\hat{\mathbf{g}}(u_0)\} = \frac{1}{2}\Delta^{-1}(u_0) \left\{ \mu_2 h_1^2 \sum_{j=1}^p g_j''(u_0)\alpha_j(u_0) - h_2^2 \beta(u_0) \right\} \{1 + o_p(1)\},$$

and the asymptotic variance is

$$\text{Cov}\{\hat{\mathbf{g}}(u_0)\} = \frac{\tilde{\nu}\nu_0}{nh_1 h_2^3 f(u_0)} \Delta^{-1}(u_0) \tilde{\Delta}(u_0) \Delta^{-1}(u_0) \{1 + o_p(1)\}.$$

□

*Proof of Theorem 3.2.* It is sufficient to show that

$$T_n = \sqrt{nh_1 h_2^3} \ell'(\boldsymbol{\theta}_0) \xrightarrow{D} N(0, T) \quad (\text{A.12})$$

where  $T = \tilde{\nu} f(u_0) \begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \tilde{\Delta}(u_0)$ , then by Slutsky's theorem and Theorem 3.1, we

can obtain

$$[\text{Cov}\{\hat{\mathbf{g}}(u_0)\}]^{-1/2}[\hat{\mathbf{g}}(u_0) - \mathbf{g}_0(u_0) - \text{bias}\{\hat{\mathbf{g}}(u_0)\}] \xrightarrow{D} N(0, I).$$

To show (A.12), we prove that for any unit vector  $d \in \mathbb{R}^{p+1}$ ,

$$\{d^T \text{Cov}(T_n)d\}^{-\frac{1}{2}}\{d^T T_n - d^T E(T_n)\} \xrightarrow{D} N(0, 1).$$

By (A.11),  $\text{Cov}(T_n) = \tilde{\nu}f(u_0)\begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \tilde{\Delta}(u_0)(1 + o_p(1))$ , and  $\text{Var}(d^T T_n d) = d^T \text{Cov}(T_n)d = \tilde{\nu}f(u_0)d^T\begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \tilde{\Delta}(u_0)d(1 + o_p(1))$ . Let  $\xi_i = \sqrt{nh_1 h_2^3} K_{h_1}(u_i - u_0) \phi'_{h_2}(y_i - \tilde{\mathbf{x}}_i^T \boldsymbol{\theta}_0) d^T \tilde{\mathbf{x}}_i$ . Similar to the proof of Theorem 2.3, we can show that  $E|\xi|^3 \rightarrow 0$ , and so the asymptotic normality for  $T_n$  holds with covariance matrix  $T$ . □

## References

- Chaudhuri, P. and Marron, J. S. (1999). Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94, 807-823.
- Chiang, C-T., Rice, J. A. and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variable. *Journal of the American Statistical Association*, 96, 605-619.
- Cohen, S.B. (2003). Design strategies and innovations in the medical expenditure panel survey. *Medical care*, 41(7):III.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.
- Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, 27, 1491-1518.
- Fan, J. and Zhang, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B*, 62, 303-322.
- Friedman J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, 9, 123-143.
- Hall, P., Minnotte, M. C., and Zhang, C. (2004). Bump hunting with non-Gaussian kernels. *Annals of Statistics*, 32, 2124-2141.
- Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society: Series B*, 55, 757-796.

- Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85, 809-822.
- Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89, 111-128.
- Kemp, G. C. R. and Santos Silva, J. M. C. (2012). Regression towards the mode. *Journal of Economics*, 170, 92-101.
- Lee, M. J. (1989). Mode Regression. *Journal of Econometrics*, 42, 337-349.
- Lee, M. J. (1992). Mode Regression. *Journal of Econometrics*, 57, 1-19.
- Lee, M.J. and Kim, H.J. (1998). Semiparametric econometric estimators for a truncated regression model: a review with an extension. *Statistica Neerlandica*, 52, 200-225.
- Muller, D. W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86, 738-746.
- Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability Applied*, 10, 186-190.
- Natarajan, S., Lipsitz, S.R., Fitzmaurice, G., Moore, C.G. and Gonin, R. (2008). Variance estimation in complex survey sampling for generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(1), 75-87.
- Ray, S. and Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *Annals of Statistics*, 33, 2042-2065.



- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, Ser. A, 26, 359-372.
- Wu, C. O., Chiang, C. T. and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying coefficient model with longitudinal data. *Journal of the American Statistical Association*, 93, 1388-1402.
- Yao, W. and Li, L. (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 1-16.
- Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labelling by highest posterior density. *Journal of American Statistical Association*, 104, 758-767.
- Yao, W., Lindsay, B. G., and Li, R. (2012). Local modal regression. *Journal of Non-parametric Statistics*, 24, 647-663.